Экспериментальный корпус белорусского языка: текущее состояние и перспективы развития

Оксана Волчек, Владислав Порицкий

Белорусский государственный университет, Минск

27 июня 2013 г.

План доклада

1. Корпусная лингвистика в Беларуси
2. Наш проект
   - Состав и структура
   - Сбор и подготовка текстов
   - Поисковый механизм
3. Примеры использования ресурса
4. Направления развития

Корпусная лингвистика в Беларуси

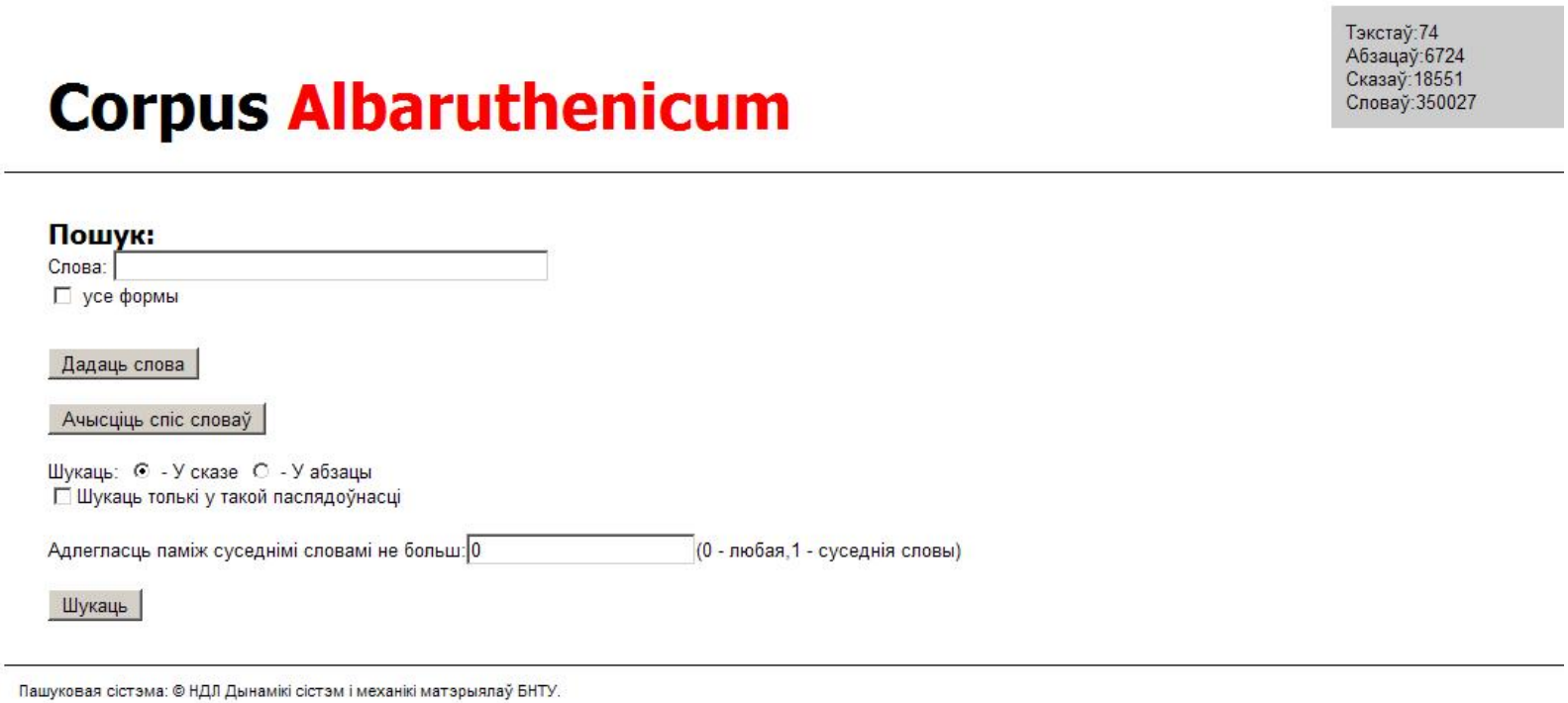
Беларускі N-корпус

- Поисковый интерфейс: `http://bnkorpus.info`
- 15 млн словоупотреблений с морфологией
- Художественные тексты разных периодов
- Опубликован фрагмент грамматической базы

Корпусная лингвистика в Беларуси

Corpus Albaruthenicum

- Поисковый интерфейс: `http://grid.bntu.by/corpus`
- 350 тыс. словоупотреблений с морфологией
- Только научные тексты
- Ограниченная функциональность поиска



Корпусная лингвистика в Беларуси

Параллельный подкорпус НКРЯ

- Поисковый интерфейс:
  `http://ruscorpora.ru/search-para-be.html`
- Русско-белорусский и белорусско-русский
- В общей сложности около 3 млн словоупотреблений с морфологией, пополняется
- Преимущественно художественные тексты

- ▷ Параллельный подкорпус НКРЯ:
  - Поисковый интерфейс: `http://ruscorpora.ru/search-para-be.html`
  - Русско-белорусский и белорусско-русский
  - В общей сложности около 3 млн словоупотреблений с морфологией, пополняется
  - Преимущественно художественные тексты

- ▷ Параллельные корпуса МГЛУ (А. В. Зубов)
- ▷ Корпус русскоязычных газет Гродненской области (Л. В. Рычкова)

Корпусная лингвистика в Беларуси

Зачем ещё один корпус?

Не хватает корпуса, который бы одновременно был:

- одноязычным;
- морфологически аннотированным;
- представительным и сбалансированный по стилям и жанрам;
- с диахронией;
- доступным для оффлайнового использования;
- свободно распространяемым.

Состав и структура проекта

1. Подкорпус газетных текстов
   - Тексты 1960-х гг.
     - «Голас Радзімы» (май 1961 – апрель 1962 гг.)
     - «Літаратура і мастацтва» (январь–декабрь 1961 гг.)

     ≈ 2 млн с/у
   - Тексты 2000-х гг.
     - «Звязда», «Чырвоная змена» (август 2008 – июль 2009 гг.)
     - «Голас Радзімы» (2008-2009 гг. полностью)

     ≈ 4.4 млн с/у
2. Подкорпус художественных текстов
   - Тексты первой половины XX века
     - Проза Якуба Коласа: «На ростанях», «Дрыгва», «Казкі жыцця»

     ≈ 0.25 млн с/у
   - Современная журнальная проза (2009-2011 гг.)
     - «Полымя», «Маладосць», «Дзеяслоў»

     ≈ 2.2 млн с/у

Состав и структура проекта

1. Подкорпус газетных текстов
   - Тексты 1960-х гг.
     - «Голас Радзімы» (май 1961 – апрель 1962 гг.)
     - «Літаратура і мастацтва» (январь–декабрь 1961 гг.)

     ≈ 2 млн с/у
   - Тексты 2000-х гг.
     - «Звязда», «Чырвоная змена» (август 2008 – июль 2009 гг.)
     - «Голас Радзімы» (2008-2009 гг. полностью)

     ≈ 4.4 млн с/у

2. Подкорпус художественных текстов
   - Тексты первой половины XX века
     - Проза Якуба Коласа: «На ростанях», «Дрыгва», «Казкі жыцця»

     ≈ 0.25 млн с/у
   - Современная журнальная проза (2009-2011 гг.)
     - «Полымя», «Маладосць», «Дзеяслоў»

     ≈ 2.2 млн с/у

Состав и структура проекта

1. Подкорпус газетных текстов
   - Тексты 1960-х гг.
     - «Голас Радзімы» (май 1961 – апрель 1962 гг.)
     - «Літаратура і мастацтва» (январь–декабрь 1961 гг.)

     ≈ 2 млн с/у
   - Тексты 2000-х гг.
     - «Звязда», «Чырвоная змена» (август 2008 – июль 2009 гг.)
     - «Голас Радзімы» (2008-2009 гг. полностью)

     ≈ 4.4 млн с/у
2. Подкорпус художественных текстов
   - Тексты первой половины XX века
     - Проза Якуба Коласа: «На ростанях», «Дрыгва», «Казкі жыцця»

     ≈ 0.25 млн с/у
   - Современная журнальная проза (2009-2011 гг.)
     - «Полымя», «Маладосць», «Дзеяслоў»

     ≈ 2.2 млн с/у

Состав и структура проекта

1. Подкорпус газетных текстов
   - Тексты 1960-х гг.
     - «Голас Радзімы» (май 1961 – апрель 1962 гг.)
     - «Літаратура і мастацтва» (январь–декабрь 1961 гг.)

     ≈ 2 млн с/у
   - Тексты 2000-х гг.
     - «Звязда», «Чырвоная змена» (август 2008 – июль 2009 гг.)
     - «Голас Радзімы» (2008-2009 гг. полностью)

     ≈ 4.4 млн с/у
2. Подкорпус художественных текстов
   - Тексты первой половины XX века
     - Проза Якуба Коласа: «На ростанях», «Дрыгва», «Казкі жыцця»

     ≈ 0.25 млн с/у
   - Современная журнальная проза (2009-2011 гг.)
     - «Полымя», «Маладосць», «Дзеяслоў»

     ≈ 2.2 млн с/у

Состав и структура проекта

1. Подкорпус газетных текстов
   - Тексты 1960-х гг.
     - «Голас Радзімы» (май 1961 – апрель 1962 гг.)
     - «Літаратура і мастацтва» (январь–декабрь 1961 гг.)

     ≈ 2 млн с/у
   - Тексты 2000-х гг.
     - «Звязда», «Чырвоная змена» (август 2008 – июль 2009 гг.)
     - «Голас Радзімы» (2008-2009 гг. полностью)

     ≈ 4.4 млн с/у
2. Подкорпус художественных текстов
   - Тексты первой половины XX века
     - Проза Якуба Коласа: «На ростанях», «Дрыгва», «Казкі жыцця»

     ≈ 0.25 млн с/у
   - Современная журнальная проза (2009-2011 гг.)
     - «Полымя», «Маладосць», «Дзеяслоў»

     ≈ 2.2 млн с/у

Состав и структура проекта

1. Подкорпус газетных текстов
   - Тексты 1960-х гг.
     - «Голас Радзімы» (май 1961 – апрель 1962 гг.)
     - «Літаратура і мастацтва» (январь–декабрь 1961 гг.)

     ≈ 2 млн с/у
   - Тексты 2000-х гг.
     - «Звязда», «Чырвоная змена» (август 2008 – июль 2009 гг.)
     - «Голас Радзімы» (2008-2009 гг. полностью)

     ≈ 4.4 млн с/у
2. Подкорпус художественных текстов
   - Тексты первой половины XX века
     - Проза Якуба Коласа: «На ростанях», «Дрыгва», «Казкі жыцця»

     ≈ 0.25 млн с/у
   - Современная журнальная проза (2009-2011 гг.)
     - «Полымя», «Маладосць», «Дзеяслоў»

     ≈ 2.2 млн с/у

Состав и структура проекта

1. Подкорпус газетных текстов
   - Тексты 1960-х гг.
     - «Голас Радзімы» (май 1961 – апрель 1962 гг.)
     - «Літаратура і мастацтва» (январь–декабрь 1961 гг.)

     ≈ 2 млн с/у
   - Тексты 2000-х гг.
     - «Звязда», «Чырвоная змена» (август 2008 – июль 2009 гг.)
     - «Голас Радзімы» (2008-2009 гг. полностью)

     ≈ 4.4 млн с/у
2. Подкорпус художественных текстов
   - Тексты первой половины XX века
     - Проза Якуба Коласа: «На ростанях», «Дрыгва», «Казкі жыцця»

     ≈ 0.25 млн с/у
   - Современная журнальная проза (2009-2011 гг.)
     - «Полымя», «Маладосць», «Дзеяслоў»

     ≈ 2.2 млн с/у

Критерии отбора журнальной прозы

(1) Рассматривались только тексты из рубрики «Проза».
Число представленных авторов – около 200.

(2) Балансировка по двум параметрам:

- Жанр
  Ориентация на жанровые пропорции подкорпуса художественной прозы XIX века НКРЯ.
- Доля произведений каждого автора
  Количество словоформ в выбранных текстах каждого прозаика $\leq 2\%$ от всей совокупности.

Демонстрационный фрагмент

https://github.com/poritski/YABC
(*Yet Another Belorussian Corpus*)

- Выборки газетных и художественных текстов в виде *txt*-файлов в кодировке ANSI
- Документация
- Доступен для скачивания вместе с поисковыми программами

Сбор и подготовка текстов

Проза Коласа, «Дзеяслоў», современные газеты

- Размещены в сети в виде HTML-страниц
- Очистка от тегов: скрипт на Perl
  - Выделяет текстовый фрагмент
  - Унифицирует передачу отдельных символов (тире, кавычки, белорусское *i*...)
  - Преобразует спецсимволы HTML (`&` , `&#945` ...)
- Идентификация дублей: скрипт на Perl
- Токенизация (здесь и далее): токенизатор Г. Шмида из проекта TreeTagger

Сбор и подготовка текстов

«Маладосць», «Полымя»

- Размещены в сети в виде PDF-файлов
- Конвертация: PDF2Text Pilot
  - Журнальная страница, свёрстанная в несколько колонок, конвертируется в одну колонку сплошного текста с группами пробелов посередине
  - Не различаются дефис и знак переноса в конце графической строки
- Поэтому осуществлялась вычитка

Сбор и подготовка текстов

Газеты 1960-х гг.

- Оцифрованы в Национальной библиотеке Беларуси как PDF-файлы без текстового слоя
- Распознавание: ABBYY FineReader 10
- Правка систематических ошибок распознавания: используется Грамматический словарь белорусского языка
  - Токены с лишним дефисом – знаком переноса
  - Самые высокочастотные неопознанные токены, содержащие явные ошибки распознавания (подстановки выписаны вручную)
  - Достаточно длинные низкочастотные неопознанные токены, которые находятся на единичном расстоянии в метрике Левенштейна от однозначно идентифицируемых словоформ из словаря
  - Подряд идущие токены, каждый из которых в отдельности не опознаётся, а результат их склейки присутствует в словаре

Сбор и подготовка текстов

Газеты 1960-х гг.

- $\approx$ 20% словоупотреблений остались неопознанными
- Но: частоты лексики близко совпадают с данными А. Е. Супруна и Н. С. Можейко по периодике 1960-70-х гг.

Поисковый механизм

К каждому из подкорпусов прилагаются две программы.

1. Индексатор:
   - Последовательно читает токенизированные файлы подкорпуса
   - Строит обратный индекс
2. Основной поисковый скрипт:
   - Читает файл *wordlist.txt* с запросами пользователя
   - Обращаясь к индексу и подкорпусу, обрабатывает запросы
   - Выводит результаты поиска в файл

Синтаксис запросов и выдача

В файле *wordlist.txt*:

- Каждый запрос записывается на одной строке
- Запрос состоит из регулярного выражения и идентификатора через знак табуляции
- Пример:

  `небасх(і|i)л(а(ў|м(і|i)?|х)?|ы|у|е)?` `небасхіл`

В файле выдачи:

- Каждому вхождению соответствует одна строка
- Приводятся информация о файле-источнике, идентификатор, точная найденная форма, правый и левый контекст (ширину устанавливает пользователь)

Примеры использования

Лексическая семантика

- Какие лексические значения развивает слово *небасхіл* (∼ рус. *небосклон*) в художественной и газетной прозе разных периодов?

`небасх(і|i)л(а(ў|м(і|i)?|х)?|ы|у|е)?` `небасхіл`

Подкорпус	Частота на 1 млн
Газеты 1960-х	4.0
Газеты 2000-х	4.8
Проза Коласа	20.2
Проза 2000-х	12.1

- Со временем семантика становится более диффузной (горизонт; часть неба над горизонтом; весь купол неба)
- Развивается или наследуется метафорическое значение: *Нясвіж – беларуская зорка на еўрапейскім небасхіле*

Примеры использования

Грамматика

- Какая из форм родительного падежа, на *-а* или на *-у*, побеждает у существительных мужского рода типа *панядзелак*?

```
(панядзелк|аўторк|чацвер)а     Род. п. -а
(панядзелк|аўторк|чацвер)у     Род. п. -у
```

Подкорпус	-а	-у
Газеты 1960-х	5	0
Газеты 2000-х	0	0
Проза Коласа	1	0
Проза 2000-х	4	0

- Грамматический словарь фиксирует только формы на *-у*, т. е. неупотребительные.

Направления развития

1. Нанесение морфологической разметки

- Существует белорусская машинная морфология (И. В. Совпель)
- В сети доступен её вариант для системы NooJ (Ю. С. Гецевич)
  - Но: низкое быстродействие; неточный набор частеречных ярлыков
- В сети доступна альтернативная лексико-грамматическая база (В. А. Кощенко и соавторы) на основе Грамматического словаря
  - Но: не целиком; опечатки, унаследованные от словаря

Направления развития

2. Совершенствование текстовой базы

- Расширение (в основном за счёт диахронии)
- Улучшение балансировки
- Доочистка газетных текстов от дублей
- Нанесение метаинформации
- Исправление опечаток

3. Доработка программной части

- Более чем однословные запросы
- Учёт грамматической информации при поиске
- Дальнейшая оптимизация быстродействия

Спасибо за внимание!